**Essd-2023-494**
ESSDD - Advancements in LUCAS Copernicus 2022: Enhancing Earth Observation with Comprehensive In-Situ Data on EU Land Cover and Use

This document is a point-by-point response to the reviews including.

## Reviewer 1 : RC1: 'Comment on essd-2023-494', Kristof van Tricht, 01 Mar 2024

***We thank the reviewer for their comments and have addressed the comments under each comment here below in bold italic.***

Manuscript summary

This paper presents the 2022 version of the Lucas Copernicus dataset. It explains the differences with the 2018 version, most notably the significant increase in number of surveyed points and the shapes of the polygons which are now significantly larger. Some basic statistics on the new dataset are also provided. The main dataset is provided as a GPKG file. The authors conclude that further harmonization is needed in order to guarantee the semantic consistency of the coding and legend, as well as the temporal inter-usability of both the 2018 and 2022 data.

Review summary

I should start by mentioning that the value of the Lucas Copernicus dataset(s) cannot be overstated. There's a tremendous amount of work and dedication that goes into the whole workflow of visiting points, interpretation of land cover, making the necessary observations and finally processing everything in a consistent polygon-based dataset. The result is one of the most influential *in situ* datasets that can be used by the Earth Observation community and an example to other countries/continents. The highly anticipated 2022 dataset will be of significant value to the community and the exciting increase in both number of observations and size of the polygons will be received with acclaim.

Next to unquestionable value of the dataset, the paper itself is generally written well. However, here and there it lacks some detail that I found essential to fully understand the nature of the dataset, required to make the bridge to EO applications. Therefore, I think some minor revisions are required to add some more detail, after which I would be happy to recommend this paper and dataset for publication in ESSD. My comments and questions for clarification can be found below.

Comments

L8-9: confusing sentence where first 82 land cover classes are mentioned, and then 88 classes. Probably different things but it could use some rephrasing for clarity.

***We thank the reviewer for their comment - indeed the readability of this sentence is wanting. We have changed the sentence to show the distinction between the LUCAS Copernicus Land cover class and the generic LUCAS Level 3 legend.***

L49: is this homogeneity interpreted also on the ortho-photo when deciding on the landcover class?

***We thank the reviewer for the comment - this is never mentioned. We include a sentence that specifically mentions the fact that the homogeneity of the window of observation and the identification of the adequate minimal mapping unit is done first via orto-photo and then confirmed or switched during the field survey.***

L62: could the authors explain the rationale behind the polygons? What is the aim of providing homogenous polygons on top of the (pure) observation of the point itself if e.g. only the respective Sentinel 10m or 20m pixel is confirmed to be homogenous? L99 might provide a clue but it would be good to explain this rationale.

We thank the reviewer for the comment - the text was lacking in clarity. We have specifically explained the rationale behind collecting the data at this distance from the point and the link to the number of Sentinel pixels that fall within he shape.

L65: this is a bit confusing as the minimum area to execute Copernicus module is reported to be 25m² while the MMU is about 79m². Where does this difference come from?

***We thank the reviewer for their due diligence. Indeed, this requirement was removed from the 2022 survey design. We have removed any mention of the MMU in the manuscript.***

L69-77: this section is not entirely clear to me. “The position they have reached” can deviate from the theoretical LUCAS point. Why and by how much? What does “cannot reach” mean? Why is a “linear feature narrower than 3m” the exception when no Copernicus-relevant information can be recorded? What are “a few meters” that a surveyor can move?

***We thank the reviewer for their comment - the text can be improved. We have added examples of the types of barriers that might prevent the surveyor from reaching the survey point. We have included information on the fact that there is no requirement to physically see the point, meaning the surveyor can be any distance away and carry out the survey via photo-interpretation. We have included a brief mention of the fact that the under 3m rule is enacted to prevent very complex landscape structures, which cannot be viewed as a pure land cover. The part about the surveyor being able to move a few meters we have removed from the text.***

L98-99: What is the aim of the quasi-circular polygon shape? Downstream applications will likely have to process the polygons further in order to be able to have e.g. pure Sentinel pixels and not a mix at the borders of the polygon where another LC could start.

***We thank the reviewer for their question. The survey protocol was designed like this initially. The 2018 data was processed without a full understanding of this protocol - something the authors have improved upon during the publishing of the current manuscript. In essence, this is the formal proper way of generating the geometries in light of how the data was collected. The reason behind this radial shape of the geometries is likely, although never mentioned in the official documents, exactly to maximise area of pure land cover.***

Sect. 5.3: where does this preliminary assessment come from? Can the authors elaborate a bit more?

***We thank the reviewer for their comment. We have expanded this section to include some specific examples and clarifications.***

L117-118: does this mean future versions of the dataset are possible? How will versioning of the dataset in that case be treated?

***We thank the reviewer for the remark. We have included a sentence stating our intention to harmonise and publish the multi-year dataset as a single unit.***

L118: Could the authors elaborate a bit more on the (planned) compatibility between 2018 and 2022 survey? e.g. what are at the moment legend inconsistencies between 2018 and 2022 and how are users advised to cope with such difference?

***We thank the reviewer for their suggestion. We have included two sentences on the matter and have provided links to publications and documents that discuss these differences in depth.***

Sect. 7: Are other dissemination methods considered in the future as well? Such as upload zo Zenodo with DOI, upload to Google Earth Engine for fast uptake by the community, …

***We thank the reviewer for the suggestion. We have included a sentence stating our intention to upload the data to GEE.***

L130-131: link only works when copy-pasting manually; hyperlink behind the text does not.

***We thank the reviewer for their observation. We have fixed all broken links, which now load the page upon clicking on link from the compiled pdf version of the manuscript.***

Figures

Figure 1: things such as the legend are too small in the current version and therefore not readable.

***We thank the reviewer for their observation. We have increased the size of the legend, which is now readable from the compiled pdf version of the manuscript.***

Figure 2: in (a), what is the size of the resulting polygon? Does this match Sentinel-1 or Sentinel-2 data? In (b), the polygon seems to contain a mix of grass, bare and builtup. Aren't the polygons supposed to contain one homogenous land cover/use type?

***We thank the reviewer for their comment. Indeed, Fig. 2b is an interesting case, because it also shows an example of when the surveyor could not physically reach the point and has done the survey from a different location, but with the idea of mapping the area around the theoretical point. This is, according to protocol, not supposed to happen, yet we believe it important to showcase examples such as this one in order to demonstrate some shortcomings of the dataset. We have included a short overview of this figure specifically in section 3.***

Figure 3: The table shown in the figure should be explained in the caption.

***We thank the reviewer for their suggestion. We have re-designed the figure and caption so that it is more self-explanatory.***

Figure 5: Why is the minimum value 0, while L65 states the Copernicus module is not executed for areas smaller than 25m²? The text on top of this figure should also be revisited. E.g. far right part is truncated. I such a large precision for these numbers required? Some rounding would probably increase readability.

***We thank the reviewer for their remark. Connected to the previous comment about L65 - we have removed any mention of the MMU. Additionally we have improved the readability of the figure by rounding the numbers.***

Figure A1: caption should contain a bit more information to be able to interpret what exactly is shown

***We thank the reviewer for their remark. We have expanded the text of the caption to include clear instructions as to how to interpret the important part of the figure.***

DATA

I checked the GPKG file and have a question after a quick look: some polygons contain hardly any data (not even landcover), e.g. point_id 38543138 has almost all "*null*" attributes while clearly being located in arable land. Why is that? In fact 2686 polygons have "*null*" in their "*survey_lc1*" attribute. Is this to be expected?

***We thank the reviewer for their due diligence. Indeed, the link was to a previous and preliminary version of the data. We have changed the file behind the link so that now the full data is available.***

## Reviewer 2 : CC1: 'Comment on essd-2023-494', Babak Ghassemi, 17 Jun 2024

***We thank the reviewer for their comments and have addressed the comments under each comment here below in bold italic.***

This paper explores the Land Use/Cover Area frame Survey (LUCAS) of the European Union, focusing on the 2022 LUCAS Copernicus module. The number of polygons increased from 60,000 in 2018 to approximately 150,000 in 2022 due to streamlined data collection protocols and refined geometry definitions. The dataset contains 88 land cover (LC) and 40 land use classes. The paper discusses the benefits of Earth Observation applications, the importance of semantic consistency, and future studies in remote sensing and computer vision, concluding with strategies for data usage and dissemination.

Here are some comments based on the existing pre-print paper as well as available data on:

https://data.jrc.ec.europa.eu/dataset/e3fe3cd0-44db-470e-8769-172a8b9e8874

Figure 3. Evaluating the provided dataset, there are 2,686 samples with null land cover attributes. Considering this issue, I would suggest either modifying the LC class count in the figure or updating the null values in the dataset.

***We thank the reviewer for their due diligence. The file behind the link was an earlier version of the data, which has subsequently been changed to the full final version, where these 2,686 missing values are present. In this light the figure is also accurate.***

Line 79 and Figure 4. There are 98 unique values in the lc1 attribute of the dataset, of which 88 are mentioned here. The following classes are missing: A3, C1, D1, D2, E1, E2, E3, F1, F2, and F3. Would it be possible to explain this difference?

***We thank the reviewer for their due diligence. The truth in the matter is that it depends which column you query. If you query column "survey_lc1" there are 98 unique values, if you query column "lc1_code" there are 88 unique values. The difference is, in a sense, not there, as exactly those classes the reviewer identified are actually just combined with their counterpart, vis-a-vis A3 == A30; C1 == C10, etc. We have harmonised these in column "lc1_code", while keeping the original data in column "survey_lc1". We agree this is a bit confusing and have included a sentence about this in section 4.***

Additionally. If this figure represents the unique labels of Level-3 LC classes, why does it contain values (whose label doesn't have 3 letters such as A, A1, B,…) possibly not

related to this category? When classifying in Level-3, can these classes be used as unique classes?

***We thank the reviewer for their comment. Indeed, some codes are not of Level 3, but these are cases where the surveyor was not sure what the specific land cover is, but filled in the field to the level they were confident. We have changed the text of the abstract and everywhere else where the number and LUCAS legend level is mentioned to reflect this nuance.***

Line 80. As mentioned in Figure 5, the mean area value is 0.35475. Therefore, it would be better for this value to be rounded to 0.35 Ha instead of 0.34 Ha.

***We thank the reviewer for their due diligence. We have corrected the typo in the manuscript.***

Section 5.2. Around 21,207 polygons out of available, 137,966 have an area below 100 m2, which means covering an area less than a sentinel-2 (or 1) pixel. For instance, pointid = 49264422, supposedly to be the apple tree has an area of around 17.38 m2, and inspecting the approximate place in Google Maps, it seems more grass or arable land. Therefore, providing an instruction regarding how to deal with this issue as well as the reliability of these polygons will be helpful.

***We thank the reviewer for their deep dive into the data. We confirm that this is the case also in the base version of the data and hence not an issue of processing, these are simply the values that the surveyor noted down. As to why they chose to give such a label to the point is a guess. We have included a few sentences in section 4, where we discuss the area of Copernicus polygons about this specific issue.***

Line 129. The dataset only contains 113 attribute columns instead of the 117 mentioned here.

***We thank the reviewer for their due diligence. With the update of the data, the number is actually 121.***

## Reviewer RC2: 'Comment on essd-2023-494', Žiga Malek, 12 Aug 2024

***We thank the reviewer for their comments and have addressed the comments under each comment here below in bold italic.***

Two other reviewers already commented on this submission, and provided a nice overview on the manuscript. I will therefore try to focus on what they might have left out, as I agree with the other two reviewers.

The authors here provide data on the most recent LUCAS polygons - improved infomation of the systematic land use survey of the European Union. Compared to the previous version, this 2022 version improved some of the limitations, and expanded on the number of points. For many smaller member states, this meant doubling or even tripling the number of points, which I am very happy to see. In general, the LUCAS polygons are a necessary companion to any remote sensing data for the European Union. In the time of AI, I am not happy to see many scientists steer away from field observations, however now more than ever, in-situ data are necessary to improve our satellite images, train environmental, agricultural and forestry models, or perform various spatial and statistical analyses. This is all necessary for the green transition and overal more sustainable ecosystem management. Therefore, I cannot overstate the importance of these data. I do not have any comments on the paper itself, it is well written and clear.

My first comment is on data availability. I am very happy to see the data already published, however also making it accessible on a cloud computing service such as Google Earth Engine or even have it on the copernicus data viewer would be super welcome. Now, the users need to download it, to see what it is about. But often, users want to check the data first, especially us that regularly work on the train, so being able to look at how useful the data are in a particular region would be great (before we for example suggest its use in a research project). Also, please provide better documentation on what is in the data. I could not find it. So, for many fields I could only guess what they are about. If these are the same as in LUCAS, fine, but still it does not harm to provide more info. It is also not in this paper, in the appendix for example. So, to facilitate use of the data, this should be easier.

***We thank the reviewer for their suggestions. We have uploaded the data to GEE under the featureCollection hosting the previous data - LUCAS Copernicus (Polygons with attributes, 2018) V1, which now holds the title - LUCAS Copernicus (Polygons with attributes) V1.***

***We have additionally added a record descriptor CSV to the repository on the JRC FTP, where one can find a breakdown of all variables and their meaning. We have also referenced this in the text.***

Secondly, the quadrilateral polygons are not well explained. You need to provide a better rationale on why you used this approach, compared to another one. Was this purely a choice given the information provided by the surveyors?

***We thank the reviewer for their comment. There seems to be a misunderstanding concerning the quadrilateral polygons - they were only generated for the 2018 data, not for the current 2022 data. The reason behind this is a misconception on our part (the authors) as to the rules, laid out by the LUCAS Copernicus module designers. We erroneously believed that the surveyed shape was the irregular quadrilateral polygon, which is created when you connect the four points in each cardinal direction. In all actuality, the surveyed polygon has a radial shape, which was true for the 2018 and 2022 survey. In the current (2022) survey, we have simply created the shapes as they were intended to be created. We have improved the text in order to make this clearer.***

You do mention the sentinel pixels - it would make sense to visualize the difference between a LUCAS point, LUCAS polygon and sentinel image, so we can imagine better. Also, when looking at the data and the paper, I am not sure I see the added value of the quadrilateral polygons instead of just having a buffer or perhaps clusters of points. I understand that it depends on the viewshed of the surveyor, but it seems that all polygons are homogeneous? Is this true?

***Indeed, we never stated clearly that the land cover information in the polygon area needs to be homogenous. We thank the reviewer for this remark. We have added the "homogeneous" epithet in s.3.***

That a polygon only covers forest for example, but not a situation where two laterals are forest, one cropland, and one grassland? I know this could be an extreme situation, but in transitional vegetation, abandoned cropland, or agroforestry areas (all land covers most difficult to map from space) this could be very useful. Can you explain a bit? I think also having data on where we see such fuzzyness could help mappers/scientists to isolate pixels where the satellite signal could no be clear. Your example on the figure, particularly Fig2a, 2b, 2d, 2f and 2h show this very well.

***We thank the reviewer for their comment. Indeed, there are caveats with both the survey design and data collection. In certain difficult cases, the produced shape might not describe the noted land cover, or the shape might be in the wrong place, or the distances noted down erroneously. We do, however, feel that having a metric of fuzziness for each polygon shape is beyond the scope of this paper and up to the community to look through their subset of the data to make sure it is clean and fit for purpose. We have specifically included examples of good and bad results in figure 2 in order to showcase these possibilities. We do, however, agree that this should be made more explicit in the text and we have improved it at the end of s.3.***

Other than that, I would like to see, in the introduction, perhaps some existing uses of the (previous) data. To demonstrate that it is being used by scientists and the wider agricultural/forestry/environment community.

***We thank the reviewer for their suggestion. While we have mentioned six separate studies, which use the 2018 data, we have augmented the introduction section with a few more relevant citations.***